NSM 2022
NORDIC STATISTICAL MEETING

Big Data for HBS –

Gains and Lessons Learned

Knut Linnerud, Statistics Norway, knl@ssb.no

Kristin Egge-Hoveid, Statistics Norway, keg@ssb.no

Abstract

Exploring and exploiting new data sources is crucial for national statistical offices (NSO) to innovate and stay relevant. To improve the quality of the upcoming household budget survey (HBS) SSB has started collecting non-survey big data from multiple commercial companies. The data contains receipt transactions from Norway's four largest grocery chains and debit card transaction data from NETS Branch Norway.

In this paper we present the recent years processes of gaining access to, receiving and exploring these new data sources. After several years of dialogue with private data owners SSB started streaming grocery receipts in real time January 1st, 2022.

New data sources often have wider areas of utilization, in both social- and business statistics. Some structural changes in the statistical office thus would be necessary. A new unit, "Team new sources" has been established, with the role of coordinating different statistical needs for new data in the office. This unit produces cross-disciplinary work: Data collection and dialogue with data owners, subject area, methods, legal expertise and IT.

We discuss challenges regarding communication and collaboration with private data owners. Scepticism and slow processes are key words. Legal issues: In challenging private companies' commercial interests. Confidentiality issues: When linking such data to persons pushes the limit of what is tolerable in the society. New data comes with a cost. And calls for a refreshed discussion of necessity and proportionality in the context of applying new data sources for the purpose of official statistics.

Keywords: Big data, Household budget survey, transaction data

1. Introduction

Exploring and exploiting new data sources is crucial for national statistical offices (NSO) to innovate and stay relevant.

The strategy for Statistics Norway (SSB) states that:

"Statistics Norway shall collect, use and share data for the benefit of society. We shall contribute to the quality and continuity of source data and exploit the growth of new data sources. We shall ensure the effective collection, use and sharing of data" (SSB 2020)

However, it is not always straight-forward to turn strategy into action.

Since 2017 Statistics Norway has had a process of gaining access to transaction data from grocery stores and banks to be used as a data source in the Household budget statistics. We have received test data, tested methods and assessed data quality for the use of transaction data in various statistics. This spring, we established and tested technical infrastructure for streaming transaction data in real time in large quantities.

However, as we write this, there is still uncertainty about access to data to produce official statistics.

The aim of this paper is to share experiences and lessons learned from Statistics Norway's work on collecting and using transaction data for the new household budget statistics.

We will focus on the prerequisites we believe are important to be able to collect and use big data in the production of official statistics, namely *the statistical act, thorough cost-benefit assessments, good internal coordination and technical infrastructure.*

We will also discuss discuss the status of the project as it stands today and some specific quality aspects applicable for transaction data.

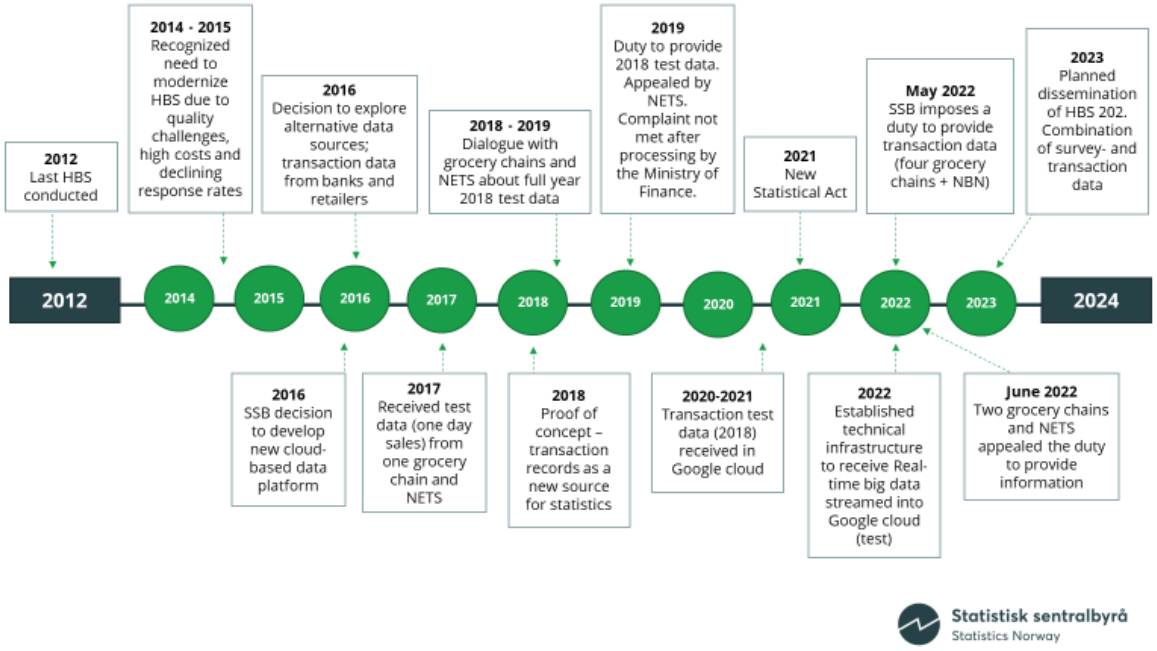
2. Background

Statistics Norway (SSB) conducted its first Household Budget Survey (HBS) in 1958. For a long period, until 2009, data were collected and published annually. The survey was last conducted in 2012.

The purpose of HBS is to provide a detailed picture of Norwegian households' annual consumption of goods and services. The household budget statistics are important to monitor changes in consumer behavior, as a weighting basis for the consumer price index (CPI) and for the audit of the national accounts. It is used to analyze the distributional effects of tax changes and to monitor development in the Norwegian diet. It is also a source for setting the size of social benefits and child support.

Until 2012, data on household consumption were collected through sample surveys only. The response burden and the survey costs where high and both sample bias and underreporting led to quality challenges. This was especially true for groceries, where the volume of goods - and thus the risk of underreporting, was greatest. Food and non-alcoholic beverages (COICOP[1] group1) are especially vulnerable for high response burden and underreporting due to the high number of items to report. There is also a requirement that food and drinks must be reported with quantity information.

Figure 1: Timeline, introduction of big data for HBS in SSB



SSB have been aiming to receive and process transaction data from the main grocery stores and Nets Branch Norway (NBN) to improve the quality of HBS-data on groceries. In 2017 we received the first test data from a grocery chain and payment

[1] Classification of Individual Consumption According to Purpose

transaction data from NBN (Fyrberg, J. et al., 2018). Based on this we concluded that these data sources had a large potential for calculating household budget shares for food. In 2020-2021 we received transaction data from three main grocery chains and NBN for all of 2018.

3. Potential of non-survey big data for HBS

SSB wish to use transaction data from grocery stores and banks to measure COICOP 01 expenditure (Food and non-alcoholic beverages) in the forthcoming HBS 2022. The plan is to combine two types of data collection for HBS: A traditional survey and transaction data for COICOP 01.

COICOP group 01 is in a special position for being investigated by transaction data because the market for groceries in Norway is dominated by four chains (with approx. 98 per cent market share). Markets for other COICOP groups are more fragmented, making it infeasible to collect receipts.

In an HBS survey, it is essential to register *what* is purchased and by *whom.* Using four separate data sources we use the following procedure to allocate receipts to households. Grocery purchases receipt data (1) are linked with payment transaction data (2), using the key variables timestamp, location and sum (Runningen Larsson, M. & Zhang, L.,2022). To further link the transactions to persons and households, a register of bank account numbers is necessary (3), and a household register (4). In addition, we need a system that automatically categorize purchases into the correct COICOP group. This will be carried out by machine learning (Jentoft, S., Toth, B. & Muller, D., 2022).

The linked data can be used to improve the quality of the HBS[2]. Either by replacing the food and non-alcoholic beverages diary component to reduce the response burden or as auxiliary information to improve the survey-based expenditure estimates.

[2] The data could also be utilized in other statistics; Consumer Price Indices, Nutrition statistics, Index of wholesale and retail sales, Business activities, Transport and tourism, Establishments, enterprises and accounts, Private health services.

TRANSACTION DATA and other data sources to be used in new HBS statistics
Bank account transactions from NETS/BAX Approx. 130 million transactions per month 2018
Receipt data from grocery stores Approx. 65 million purchases per month 2018 Approx. 650 million sold items per month
Register of bank accounts (tax directorate) 33 mill. observations/accounts in 2018
Household register (based on population register)
COICOP training data – Consumer price index

By combining these sources, we could achieve the following advantages:

- Data that cover the whole year
- Include close to the whole target population
- Improved quality and timeliness
- Lowered response burden
- Reduced internal cost compared to surveys
- Potential to further combine with both register data and survey data

Receipt data from grocery chains

Receipt data from the four dominating chains (NorgesGruppen, Coop, Rema1000 and Bunnpris) in the grocery market have great potential for producing HBS statistics. The transactions include almost 98 per cent of all purchases of groceries[3]. The receipt data contain key variables which enable us linking them to payment transaction data; Timestamp, shop and sum. They also contain bar code (GTIN nr), descriptive text, amount sold, and price.

[3] The receipt transaction data will not cover grocery purchases carried out from other shop types - like online purchases, independent stores, marketplaces - and purchases abroad. These are together estimated to constitute 2 per cent of grocery purchases in Norway.

Payment transaction data

Payment transaction data from Nets Branch Norway includes all Bank Axept payments via debit cards. They contain key variables which enable us to link to receipt data; timestamp, shop and sum. They also include information about method of payment (if items are paid partially by cash/other cards, or if cash is withdrawn), and possible returns. All variables important in the statistical production to minimize sources of error (Runningen Larsson, M. & Zhang, L.,2022).

The share of payments carried out by debit cards constitute approx. 75 per cent of all payments in detail trade. Coverage is very good, but we exclude grocery purchases done by cash and/or credit cards. Purchases by Apple Pay, Google Pay or customer clubs could not be identified in the payment transaction data either.

Table 2 Payment methods, grocery purchases in Norway, 2018

Payment methods	**Per cent**
Cash	<3
Credit card	15-20
Debit card	75
Other (E.g Apple Pay, Google Pay)	2-5

SSB has sent a duty to the grocery chains and NBN to provide all transactions for a test period for two years (in total, approx. 1,3 billion receipts/1,6 billion payment transactions per year). During these two years we will test possibilities to produce the statistics with considerable smaller data amounts, either by sample of shops, sample of days/weeks, and/or sample of persons.

4. Important prerequisites to access privately held big data for official statistics

Accessing and using big data in the production of official statistics is an innovative step forward. Most statistical offices do not yet have much experience in collecting and using this type of data to produce statistics.

Unlike public administrative registers, the new big data sources are special in the sense that they are often owned by private companies that are governed by the market and profitability. These are data owners to whom Statistics Norway traditionally does not have the same close relations when it comes to data exchange, as those who manage public registers. Nor are they subject to the same state control as public actors.

The data these companies possess is often sold or used for their own profit. In our case, the data also includes third-party actors, customers, who cannot be expected to be aware that Statistics Norway will use the electronic tracks from their purchases and payments to produce statistics. There are therefore many considerations to take when collecting this type of data.

We have identified what we consider to be four important prerequisites:

1) The Statistics Act
2) a thorough cost-benefit assessment
3) internal coordination
4) technical infrastructure to receive and process big data

1. The Statistics Act

In 2021, a revised and new Statistics Act for Norway came into force. The purpose of the Act is *"to promote the development, production and dissemination of official statistics with a view to increasing public knowledge, and providing a basis for analysis, research, decision-making, and general discussion in society" (SSB 2021).*

The Act authorizes Statistics Norway to impose a *duty to provide information* for any data, public or private, that may be necessary, to develop, produce or disseminate official statistics.

In the preparatory work for the new Statistics Act, special emphasis was placed on adapting the new Act to also reflect the ever-increasing amount of data in society and the need to utilize new types of data to constantly improve the quality and timeliness of official statistics.

The Statistics Act and Statistics Norway's legal basis to impose a duty to provide information has been an important and fundamental precondition in the process for gaining access to new big data sources.

2. Cost-benefit assessment

The Statistics Act is a powerful tool for gaining access to new data sources. However, before the duty to provide information can be imposed, Statistics Norway must conduct a thorough cost-benefit assessment of the planned data collection. The assessment must be made publicly available (§ 10 (5) The Statistical Act)

The assessment must include:

- the purpose of collecting the data
- justification of the benefits of collecting the data
- the cost for those required to provide information, including privacy costs for individuals
- assessment of the sensitivity of the information
- description of special security measures, if there is a need for such
- justification for the scope of data required (data minimization)

The costs and benefits must be weighed and be proportionate. The duty to provide information must be carried out in a manner that imposes the lowest possible burden on the involved parties.

The cost-benefit assessment made for the transaction data is important for several reasons. Most importantly it ensures a thorough process internally in NSO that the decision on the duty to provide information is made on the correct basis and within the mandate of the Statistics Act and the Personal Data Act.

In addition, the publication of the cost-benefit assessments provides openness and transparency about the assessments made by Statistics Norway. This enables an open debate on the utilization of such data to produce official statistics.

In order to make these assessments, there has been a need for close collaboration internally between the legal department, statistics department, Methodology, IT and management.

Privacy assessment

Receipt data are not personally identifiable in themselves. Only after receipts are linked to card transactions they are considered personal identifiable information. Debit card transactions are personally identifiable.

The transaction data is special both due to the large amount of data and the fact that the information does not already exist in public registers. The data is received in close to real time and with a high degree of detail. The data will be linked to the identified person and the household, together with information such as income and education. The data will contain information about where and when you have bought groceries, as well as detailed information about which goods and quantities of goods you have purchased.

An important starting point for Statistics Norway's work with this type of data has been that the individual consumer cannot be expected to be aware that Statistics Norway will use the electronic tracks from their current purchases, and further connect these with personally identifiable data to create statistics.

To ensure privacy, Statistics Norway has general security measures that apply to all processing of statistical information in Statistics Norway. The Statistics Act requires us to ensure confidentiality in all dissemination of statistics, and to implement measures to achieve a satisfactory level of security.

Further, the information can only be used for statistical purposes within the framework of The Statistics Act, which constitutes a guarantee that this information cannot be used for control purposes or other purposes of direct significance to the registered person. Statistical use is, on a general basis, considered to be of low privacy risk.

However, due to the special nature of this data, the collection and use of the data is believed to be perceived as more intrusive than the utilization of administrative register data. Statistics Norway has therefore initially adopted a duty to provide information for a full-scale collection of transaction data for a two-year period, until 2024. In this period, we will seek to implement measures to protect privacy and safeguard data security, in addition to the general protection measures already applied to administrative register data. We will continuously work to develop methods

for data minimization, such as indirect links, aggregations and sampling, which will reduce the privacy consequences. After the end of the two-year period, we must make a new assessment of the scope of data collection.

Reactions from data owners after submitting a duty to provide data in May 2022

The decision on the duty to report transaction data (authorized by the Statistical Act) has been appealed by two out of four grocery chains, and by NETS Branch Norway. Privacy concerns and data security are the main objections, in addition to requirement for data minimization. Media has also shown interest and the Norwegian Data Protection Authority has been involved. We can probably expect a principal discussion of necessity and proportionality in the context of applying these data sources for purpose of official statistics. When the conference is held in Reykjavik in August, the appeals from the data owners have not yet been processed.

SSB believes the collection of transaction data is legal under the Statistics Act and the Privacy Act, and that we can argue for the need / benefit of a comprehensive collection of data for a limited period. We also believe that we meet the necessary measures to ensure privacy and data security. Still, we must acknowledge the concerns and responsibility private data owners have, to ensure privacy and data security for their customers. Regardless of the NSIs ability to secure data and protect privacy, many individuals will still have concerns about the use of their private data. The data owners concern is that they could expect customers questioning their data security in general, jeopardising the company's position in the market.

These concerns should be seriously met by the NSI, and we regard it especially necessary to have a communication strategy that could meet the experienced uncertainty regarding both data security and privacy protection concerns.

3. Internal communication and coordination

When an NSO decide to start a process of getting access to new data sources for new statistics or to improve existing statistics, it is important that the whole organization pull together in the same direction.

When considering new data sources, several questions arise:

- Are similar types of data already available in the NSO?
- What is the potential of the specific data source? Can the data source be applied in different statistics?
- Have any other unit in the NSO been in contact with the data owner?
- How can we approach the data owner?

An important experience gained in Statistics Norway is the need for good internal coordination with clear role clarification and division of responsibilities. Both statistical needs, legal assessments and the technical scope must be assessed before taking the initiative to obtain new big data sources. In addition, there is a need to build a relationship with data owners.

With many different stakeholders, initiatives and opinions, ineffective lines of communication can easily arise, something we experienced at Statistics Norway.

Hence, in SSB a “Team New Sources” has been established, responsible for coordinated advances against private companies’ new data sources. They produce overview of relevant users in the NSO for different types of new data sources, and over which sources are already in use. It is especially important to avoid that different NSO units contact data owners with similar or close to similar data demands.

The Team New Sources assist with the following:

- Coordinate NSO initiatives for data access
- Contact and dialogue with data owners
- Assist statistical units and the IT department in dialogue with data owners about technical solutions, follow-up and necessary support.
- Build trust with the data owner

In the process of identifying NSO units that could benefit a new data source, it is important to carefully consider the size of the initiative. Many potential users mean a high benefit in the NSI. At the same time, many potential users mean the process will necessarily slow down (more internal coordination, and compromising content of delivery, aggregation level, frequency and technical solutions for storing and editing

the data)[4]. A delayed process is undesirable, since it challenges planned dissemination of statistics. And necessary time for testing and quality checks of the data would be stressed.

Summarized: For statisticians it is good help to have a “professional” team to assist in the processes. We believe that, with a clear division of responsibility, the processes against data owners are better protected, and resource saving.

4. Technical infrastructure - what is needed to utilize big data?

The large amount of data from the new sources (receipts from grocery stores, bank transactions) requires new systems for receiving, storing and editing data. It is estimated that SSB will receive 1.3 billion receipts and 1.6 billion payment transactions per year[5]. This makes it challenging to carry out statistical production in an environment based on SSB’s existing “on-premise” setup.

A strongly connected and secure infrastructure is needed. SSB has developed a new cloud-based services/platform as a part of a bigger modernization process in the office. The new data platform manages all types of data that Statistics Norway collect, treat and disseminate as statistical products. Google Cloud Platform (GCP) is the service provider, and the new platform in cloud is called DAPLA (DAta PLAtform).

DAPLA includes a computing environment that enables fast data processing and analysis, using SPARK cluster to process unlimited data in parallel. The new tool landscape includes tools such as R, Python and Jupyter notebooks. All processes and solutions support securing personal identifying information. All data on the platform is encrypted and all personal identifying information are pseudonymized.

[4] Demand of data content could e.g. be very different between e.g. social statistics (detailed data) and business statistics (more aggregated data).

[5] In 2022-2023 we will test possibilities to produce the statistics with considerable smaller data amounts, either by sample of shops, sample of days/weeks, and/or sample of persons.

Continuous streaming of real-time data

After several years of dialogue with private data owners SSB started streaming grocery receipts in real time, from the four main grocery chains, from January 1st, 2022. Utilisation of the data is per June 2022 put on hold, pending determination of the appeals from data owners of the duty to deliver data.

Continuous streaming of data provides advantages in form of a more stable delivery, improved timeliness and thus possibilities for developing new innovative statistical products.

Set-up of technological solution was carried out in a close cooperation between SSBs IT department and the IT units of the private companies. Many others were involved in the process; the legal unit, HBS statisticians and the data collection unit. Real-time data is streamed into the new cloud platform and made available for users in JupyterHub.

Challenges experienced related to big data on the cloud-based platform[6]

- When receiving big data, data quality check at an early stage is essential. If errors are discovered after transferring and encrypting the full dataset, you need to start all over, and weeks/months are lost.
- Data processing in the cloud-based data platform requires high level data scientist skills in new tools e.g. Python and R. This kind of personnel are highly sought and challenging to recruit in Statistics Norway.
- The development of the new cloud-based platform simultaneously with attending transaction data for the HBS challenged progress for the project.
- When the HBS entered the new cloud-based data platform all parts of the HBS statistical production had to be re-established.

[6] Due to the late process of accessing 2022 retail transaction data real-time streaming, we have few experiences to share regarding quality of data, and experiences with real-time streaming itself.

5 Quality aspects for new data sources

The use of new data sources like transaction data has many advantages. However, the access to new data sources also challenges the existing data quality standards in the NSO. Can we achieve acceptable statistical quality standard for the new data types? In this chapter we will describe different quality aspects especially valid for the new data sources used in HBS.

Data quality issues in general

To which extent is the availability and quality of an external data set guaranteed over time? NSOs have only limited control over the availability and quality of an external data source. There will always be a risk that a data source will be unavailable in the future, in a different issue or in different quality. Data sources could change structure or go out of use. This should be acknowledged before considering a data source for official statistics.

It is essential to have good dialogue and cooperation with the data owners. This enables statisticians and provides a meeting point for discussing any issues about the delivery, such as variables, metadata and the representativity for the data. If there is lack of a fruitful dialogue with the provider, the quality control might be poor.

A formalized cooperation about quality between NSO and the data owner is recommended, for example in form of a written contract including obligation for data owner to contact NSO about any possible changes in the data stream.

Processing errors

Processing errors can occur using transaction data. It is recommendable to invest in a monitoring system that checks data quality. E.g. for counting the number of valid cases in the streaming data, and setup of an automatic notification/report if data flow stops for shorter or longer period. Processing errors must preferably be detected at an early stage. When considering the extreme amounts of transaction data, it's even more important to detect abnormalities at an early stage, since re-transmission would be especially time-consuming. Real-time streaming of data is considered to reduce processing errors.

Representativity

The market for groceries in Norway is unique, with only four large chains representing approx. 98 per cent of the market (described in chapter 2). Of all receipts, debit card transactions (which are the ones we receive) constitute almost 75 per cent. By linking grocery receipts to bank transactions and bank accounts, we could connect approximately 70-75 per cent of the receipts to persons and households (Runningen Larsson, M. & Zhang, L.,2022). Purchases made by credit cards, cash, via customer clubs and other payment methods can't be linked.

There is uncertainty connected to the future representativity of debit card transactions. Payment methods are complicated and evolving fast. New constellations and payment methods are continuously introduced to the market. E.g. Apple pay, Google pay and customer cards. These are all payment methods that can't be tracked by receiving debit card bank transaction data. Despite the huge amounts of transaction data received, they still can't fully cover the target population. There are population groups that particularly will have under-coverage using the new data types (e.g. elderly people paying by cash). This is an important issue to take into account when using big data for official statistics.

Coherence and comparability

Another important issue using new big data, is how we can secure standardization of data sources and methods across countries, and over time. The use of new data sources and methods can compromise this type of comparability, which is undesirable. Necessarily, this would imply introduction of new methods, and comparability is challenged. For example, to meet strict requirements of privacy protection, a precaution could be to aggregate observations for groups of households. Traditional microdata then couldn't be obtained, with the implication of possible disruptions of time-series.

Time aspect

Slow processes (and uncertainty) of getting access to data, transferring and storing data in cloud challenges planned dissemination of statistics.

This is the most imminent quality aspect. In an innovative phase, necessary time for set-up, testing, quality checks and dissemination is stressed because of delayed data access.

6 Summary

In this article we have examined the last years experiences with transaction data for HBS in Statistics Norway.

Collecting transaction data is the first step Statistics Norway takes in collecting big data in real time. The data is very detailed and must be treated as highly sensitive, both in its raw and linked form. It is essential for the NSO and the data providers that both commercial interest and data confidentiality are protected.

We have tried to carefully assess and discuss the technical benefits of efficient and frequent data collection and the opportunities this provides in a long-term perspective up against privacy consequences, what The Statistics Act gives us a mandate for, and the significance of collecting such data for society's trust in Statistics Norway. This has been a necessary and important discussion, and a fundament for further work.

Slow processes are key words. One lesson learnt is that dialogue with private data owners could suddenly stop, at any stage in the process. There’s ongoing risk for non-cooperation and late deliveries from the private data owners, which jeopardize the production process and quality aspect. Still, in June 2022, data access can’t be taken for granted. Time will show if we are able to disseminate a new HBS in 2023 based partially on new big data sources. There’s ongoing and future work to develop methods for data minimization, such as indirect links, aggregations and sampling, which will reduce the privacy consequences.

Milestones gained, HBS project:

- Test transaction data for 2018 received, stored and linked to persons/households.
- Consumption 2018 for COICOP 01 calculated and evaluated. Very good results.
- Established project in GCP enables adaptation of unlimited amounts of data.

- Established streaming real- time big data to the new data platform
- Skills in new tools (Python, R) gained.
- Established A/I categorization of products using machine learning to predict COICOP five-digit level.
- The production system established for test data ready to be re-used for new volumes.

References

Runningen Larsson, M. & Zhang, L.,2022. *Using non-survey big data to improve the quality of the household budget survey*

Jentoft, S., Toth, B. & Muller, D., 2022. *From manual to machine: Challenges in machine learning for COICOP coding*

Fyrberg, J. et al., 2018 *Proof of concept – transaction records as a new source for statistics, Oslo: Statistics Norway Internal Notes*

Statistics Norway (2020): *Strategy for Statistics Norway, Plans and reports 2020/7*

Statistics Norway (2022): *Cost-benefit assessment pursuant to The Statistics Act § 10 (5)*: https://www.ssb.no/omssb/ssbs-virksomhet/kost-nyttevurdering/leveranse-av-bongdata-fra-dagligvarekjedene-rema-1000-norgesgruppen-coop-og-bunnpris

Statistics Norway (2021): Act relating to official statistics and Statistics Norway (the Statistics Act)